Chronopolis, University of North Texas, MetaArchive: PRESERVATION COOPERATION

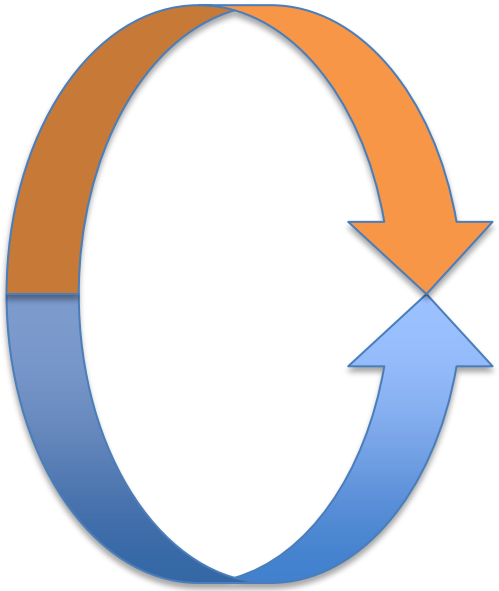
David Minor, SDSC/UCSD

Mark Phillips, University of North Texas

Matt Schultz, MetaArchive

Basic Question

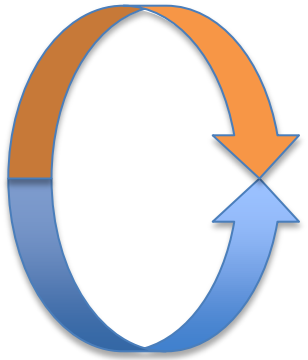
How can preservation systems share objects?



Basic Question

How can preservation systems share objects?

- Different infrastructures
- Different operating procedures
- Same broad outlooks on preservation



After several years of discussion, an NHPRC grant provided the opportunity to undertake the work.

The Teams

Chronopolis

MetaArchive

University of North Texas

Chronopolis

- A digital preservation network developed by a national consortium, with initial funding from NDIIPP
- Based on SRB/iRODS with additional tools layered on top
- Has a current storage capacity of 300 TB (100 TB at 3 nodes)
- Has geographically distributed copies of all data
- Has detailed monitoring and auditing

MetaArchive

- Established in 2004 (support from NDIIPP and NHPRC), preserving content for 16 members
- Uses LOCKSS software to provide peer-to-peer distributed digital preservation infrastructure
  - All content is stored in multiple copies at geographically dispersed locations.
- Sustainable organizational framework: Membership organization with a 501c3 host (Educopia Institute)
- 254 TB network capacity (and growing)

University of North Texas

- Digital repository infrastructure built with Curation Micro-Services
- Archival management and end user content delivery systems
- Generic workflow for packaging, transferring and ingesting content
- 45 TB of data archived with a 100 TB local capacity
- Technology neutral replication system

Stage One: Basic transfers

Stage One: Basic transfers

From MetaArchive LOCKSS-based system into Chronopolis' SRB-based system, through UNT

Two different transfer approaches:

- BagIt
- SRB client

BagIt

BagIt is a hierarchical file packaging format for the exchange of generalized digital content.

- There is no software to install
- Consists of base directory with manifest file & subdirectory with content
- Manifest file has a row for each content file with:
  - Full path in content directory
  - A checksum for file

Holey Bags

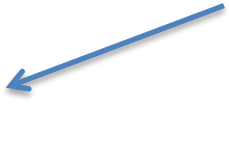
- Have additional 'fetch.txt' file in base directory & empty content directory
- URLs for each content file are listed in fetch.txt file.
- Can reduce transfer time by fetching content in parallel

http://www.digitalpreservation.gov/library/resources/tools/docs/bagitspec.pdf

Simple BagIt transfer

- Transfer of 200MB archival units
- Checksum-based verification
- Registration into MCAT

SRB/iRODS client transfer

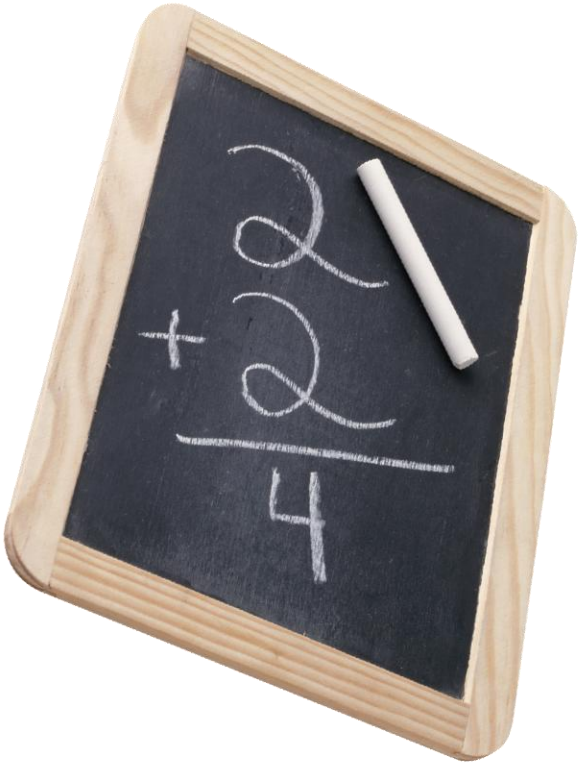


- Script to create bags

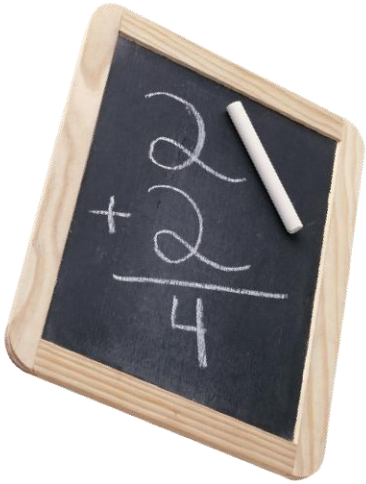
- Script to “put” bags using SRB client

- Registration into MCAT

Lessons learned from stage one



Lessons learned from stage one



- Authentication and registration issues
- Issues with character consistency
- MetaArchive AUs must be taken out of active preservation mode and be rendered static before being Bagged, otherwise the LOCKSS re-crawling and polling/voting processes will interfere with their packaging.

Stage one made us want to …

- Measure transfer rates to determine if one method is more efficient or provides better service
- Compare usability of SRB/iRODS client transfers with BagIt wgets through standard web channels
- Transfer collections in excess of 1TB to test large-scale efficiency of methods

Stage two: improved transfers

Using a 1.5 TB collection from Folger Shakespeare Library

Stage two: improved transfers

- UNT configured a 50TB server on-site as a MetaArchive-LOCKSS cache, populated with a proxy export from MetaArchive member GA Tech
- UNT's cache participated a full round of LOCKSS-driven file voting/polling validation and ensured 100% integrity of Folger collection content
- UNT developed a custom script that exploits the built-in LOCKSS content serving features and relies upon open source micro-services to retrieve and validate the Folger files, and package each archival unit according to the BagIt specification

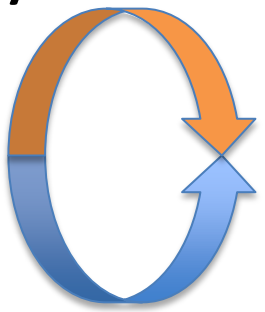
Stage two: improved transfers

- Chronopolis provided a script to crawl BagIt manifests, verify checksums, check inventory synchronization and account for all Bags transferred
- Preliminary transfer rates were tested on a 6GB archival unit subset of Folger collection content and the 1.3 TB was transferred over the course of a 48 hour period
- Chronopolis ingested 250 Bags
- UNT, Chronopolis and MetaArchive staff began evaluating requirements for ensuring that the Conspectus data management tool and its associated collection level metadata could be exported into the Chronopolis environment.

Stage three (in process)

- Transfer content from Chronopolis to MetaArchive
  - Aggregations for non-MetaArchive content
  - Loading data into MetaArchive from Bags

- Transfer content from UNT's repository to Chronopolis and MetaArchive



Conclusions

Conclusions

- Simple micro-services approach to interoperability a huge success
  - BagIt, common code/tools, iRODS commands
- Enhancements to features and workflows
  - LOCKSS Export API & Chronopolis auto-validate
- Need to explore common data management approaches between MetaArchive, Chronopolis, and UNT

People

- Chronopolis:
  - David Minor – minor@sdsc.edu
  - Don Sutton – suttond@sdsc.edu
- MetaArchive:
  - Matt Schultz – matt.schultz@metaarchive.org
- University of North Texas:
  - Mark Phillips – mark.phillips@unt.edu
  - Kurt Nordstrom – kurt.nordstrom@unt.edu